IFM 5.14 Künstliche Intelligenz, W19

Blatt 6: Entscheidungsbäume (22 Punkte)

Praktikum: 27.11.2019

Carsten Gips, FH Bielefeld

1 Handsimulation CAL2 (4 Punkte)

Zeigen Sie mit einer Handsimulation, wie CAL2 mit dem folgenden Trainingsdatensatz schrittweise einen Entscheidungsbaum generiert. Nutzen Sie die linearisierte Schreibweise.

Beispiel	x_1	x_2	x_3	Klasse
1	a	a	a	1
2	a	b	a	2
3	a	a	b	1
4	b	a	b	1
5	a	a	c	1
6	b	b	b	2

Thema: Anwendung von CAL2

2 CAL2: Anforderungen an Trainingsmenge (3 Punkte)

Welchen Entscheidungsbaum würde CAL2 lernen, wenn dem Trainingsdatensatz aus der vorigen Aufgabe der Vektor $((a, a, b), 2)$ als Beispiel Nr. 7 hinzugefügt werden würde?

Thema: Voraussetzungen an Trainingsdatensatz

3 Pruning (3 Punkte)

Vereinfachen Sie schrittweise den Baum $x_3(x_2(x_1(C, A),\ x_1(B, A)),\ \ x_1(x_2(C, B),\ A))$ so weit wie möglich. Nutzen Sie die linearisierte Schreibweise. Geben Sie die jeweils verwendete Regel an.

Thema: Anwendung der Transformations- und Pruning-Regeln

4 Handsimulationen CAL3 und ID3 (4 Punkte)

Es ist wieder Wahlkampf: Zwei Kandidaten O und M bewerben sich um die Kanzlerschaft. Die folgende Tabelle zeigt die Präferenzen von sieben Wählern. Führen Sie je eine Handsimulation mit CAL3 ($S_1 = 4$, $S_2 = 0.8$) und ID3 durch.

Nr.	Alter	Einkommen	Bildung	Kandidat
1	≥ 35	hoch	Abitur	O
2	< 35	niedrig	Master	O
3	≥ 35	hoch	Bachelor	M
4	≥ 35	niedrig	Abitur	M
5	≥ 35	hoch	Master	O
6	< 35	hoch	Bachelor	O
7	< 35	niedrig	Abitur	M

Thema: Verständnis algorithmischer Ablauf CAL3 und ID3

5 Machine Learning mit Weka (6 Punkte)

Weka[1] ist eine beliebte Sammlung von (in Java implementierten) Algorithmen aus dem Bereich des Maschinellen Lernens. Laden Sie sich das Tool in der aktuellen stabilen Version[2] herunter und machen Sie sich mit der beiliegenden Dokumentation vertraut.

Laden Sie sich die Beispieldatensätze „Zoo“ (`zoo.csv`) und „Restaurant“ (`restaurant.csv`) aus dem AIMA-Repository[3] herunter.

Lösen Sie nun folgende Aufgaben:

a) Zum Laden der Beispieldatensätze in Weka müssen die `.csv`-Dateien eine Kopfzeile mit den Namen der Attribute haben. Passen Sie die Dateien entsprechend an und laden Sie diese im Reiter „Pre-Process“ mit „Open file . . . “.

b) Wechseln Sie dann auf den Reiter „Classify“ und wählen Sie mit dem Button „Choose“ den Entscheidungsbaum-Lerner J48 aus.[4] Lernen Sie für die beiden Datensätze je einen Entscheidungsbaum. Wie sehen die Bäume aus? Wie hoch ist jeweils die Fehlerrate für den Trainingssatz?[5] Interpretieren Sie die Confusion Matrix.

c) Lesen Sie in der beiliegenden Doku zum Thema „ARFF“ nach. Dabei handelt es sich um ein spezielles Datenformat, womit man Weka mitteilen kann, welche Attribute es gibt und welchen Typ diese haben und welche Werte auftreten dürfen. Erklären Sie die Unterschiede zwischen „nominal“, „ordinal“ (bzw. „numeric“) und „string“. Konvertieren Sie den Zoo- und Restaurantdatensatz in das ARFF-Format. Beachten Sie, dass die ID3-Implementierung von Weka nicht mit bestimmten Attributtypen umgehen kann.

d) Lernen Sie für die im ARFF-Format vorliegenden Datensätze (Zoo und Restaurant) erneut Entscheidungsbäume. Nutzen Sie diesmal sowohl ID3 als auch J48. Vergleichen Sie wieder die Ergebnisse (Entscheidungsbäume, Fehlerraten, Confusion Matrix) untereinander und mit den Ergebnissen aus dem J48-Lauf mit den `.csv`-Dateien.[6]

Thema: Kennenlernen von Weka

6 Anwendungen (2 Punkte)

Recherchieren Sie, in welchen Anwendungen Entscheidungsbäume eingesetzt werden. Erklären Sie kurz, wie die Bäume jeweils trainiert und wofür sie genutzt werden.

Thema: Anwendungen von Entscheidungsbäumen

[1] cs.waikato.ac.nz/ml/weka

[2] Wenn Sie *Weka 3.6* einsetzen, sind alle für dieses Blatt erforderlichen Algorithmen bereits vorhanden. In neueren Versionen müssen Sie in der Weka-Haupt-GUI den Paketmanager unter „Tools“ starten und dort nach einem Paket suchen, welches ID3 enthält, und dieses Paket nachinstallieren.

[3] github.com/aimacode/aima-data

[4] Dies ist eine Java-Implementierung von C4.5. Die ID3-Implementierung funktioniert für den `zoo.csv`-Datensatz leider nicht . . .

[5] Stellen Sie unter „Test options“ den Haken auf „Use training set“.

[6] Entfernen Sie für ID3 im geladenen Zoo-Datensatz das erste Attribut („name“): In der Pre-Process-Ansicht das Attribut auswählen und den Button „Remove“ drücken.